DOCUMENT RESUME

ED 163 072 TM 008 154

AUTHOR Ironson, Gail H.
TITLE A Comparative Analysis of Several Methods of Assessing Item Bias.
PUB DATE Mar 78
NOTE 54p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978)

EDRS PRICE MF-$0.83 HC-$3.50 Plus Postage.
DESCRIPTORS Aptitude Tests; Comparative Statistics; Complexity Level; Culture Free Tests; *Item Analysis; *Racial Discrimination; Senior High Schools; *Test Bias; *Test Items; Test Validity
IDENTIFIERS Chi Square; Item Characteristic Curve Theory; Item Discrimination (Tests)

ABSTRACT

Four statistical methods for identifying biased test items were used with data from two ethnic groups (1,691 black and 1,794 white high school seniors). The data were responses to 150 items in five subtests including two traditional tests (reading and mathematics) and three nontraditional tests (picture number test of associative memory, letter groups-inductive reasoning, and mosaic comparisons-perceptual speed). The four statistical methods compared indices for the two groups with respect to: a normal transformation of percent passing, the point biserial item test correlation, chi square, and the area between item characteristic curves (ICC). The results for three of the methods were moderately positively correlated, but there was very little agreement between the item test correlation method and the other three. A random sample of 100 black students was matched with a sample of 100 white students on the basis of total test score, and item data were compared for percent passing in the two groups. Finally, a comparison was made of the indications of bias in traditional tests and the indications of bias in the nontraditional tests. (CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Session 16.24

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED163072

A Comparative Analysis of Several Methods of Assessing Item Bias*

by

Gail H. Ironson

Bowling Green State University

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Gail H. Ironson

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Presented at the

AERA Annual Convention

Toronto, March 1978

*Sincere thanks are expressed to Dr. Michael Subkoviak, for his advice through this project and to Dr. Hunter Breland

IM008 154

Controversy over testing practices has been with us ever since the inception of wide scale testing during World War I (Cronbach, 1975). More recently the issue of bias in measurement and selection has been a growing concern. Minority groups have claimed that traditional educational and employment tests are oriented toward a white middle class culture, and may not reflect the true ability of minorities (Williams, 1971, 1970). Since a majority of the people on whom the tests are developed are white, those items functioning best for them are retained. However, Green and Draper (1972) showed that those items which are psychometrically best for whites were not necessarily the same as those chosen as best for blacks by various item analysis procedures. One might suspect the same would be true for poor whites or other cultural groups.

Fairness Versus Bias

Before proceeding to a delineation of statistical models of bias, a distinction between the terms fairness and bias should be made. Often the two terms are used interchangeably as in Green and Draper (1972):

> A biased test is popularly understood to be a test which is unfair to identifiable subgroups of the general population in which it is being used. (p. 1)

On the other hand, Angoff (1976) and Rudner (1976) use an example to emphasize the point that fairness and bias should be considered as distinct. Consider the test item: "Name three parks in Manhattan." This item may be statistically biased in the sense

that it is relatively more difficult for people who do not live in New York City than those who do. Yet this item might be fair in the sense of being valid for the purpose of selecting New York City taxicab drivers. Thus, both Angoff (1976) and Thorndike (1971) believe a judgement of fairness must be based upon an examination of the purpose and intended use of the test. Angoff (1976) states:

> If a test is prospective in its purpose, fairness should be evaluated with respect to its ability to predict future performance for the groups of individuals. A retrospective test which is intended to evaluate the assimilation of information should be evaluated in terms of opportunities for equivalent past exposure (p. 3).

In this paper, bias will refer to a statistical inequity in the way a test or item is functioning in different groups. The view is taken, in agreement with Darlington (1976, p. 43) that test fairness "must be handled rationally--that is by human judgment--rather than mechanical formulas."

Internal Versus External Criterion

A distinction between two approaches to assessing bias--in the presence of an external criterion versus in the absence of an external criterion--was first made by Pothoff (1966). This distinction is the basis of two major lines of research in the area.

The first, called "bias in selection," arises when a test is used for hiring or screening purposes because the test correlates with an external criterion of success. For example, the Scholastic Aptitude Test is used to screen high school students for college

because of its relation to grade point average (GPA). The question of interest is whether such selection procedures deal justly with both minority and majority groups. The present paper does not deal with the bias in selection, but the interested reader should refer to the special issue of the Journal of Educational Measurement (Spring, 1976) which is devoted entirely to this topic.

The second line of research is concerned with detecting biased test items in the absence of an external criterion. Bias in this sense is often judged by examining the statistical properties of an item in relation to the other items of the test. Definitions of what a biased item is have included: (1) an item that measures different traits across culture groups (Green and Draper, 1972), (2) an item which behaves differently for members of different groups (Merz, 1976), (3) an item which is relatively more difficult for one group as compared to a second group and (4) an item which has different statistical characteristics for equal ability members of the two different groups.

Proposed methods for measuring item bias have proliferated in the past two years and have been reviewed by Merz (1977) and Rudner (1976, 1977c). Briefly, these methods include techniques that examine relative item difficulty across different groups or item by group interaction (Angoff & Ford, 1973; Cleary & Hilton, 1968; Echternacht, 1974), differences in the point biserial discrimination index across groups (Green & Draper, 1972; Ozenne, Van Gelder, & Cohen, 1974), differences in the item characteristic curves for different groups (Lord, 1976; Scheuneman, 1975, 1976),

differences in the distribution of incorrect responses for various groups (Veale and Foreman, 1976) and finally, differences in multivariate factor structures across groups (Green [1976], Green and Draper [1972], and Merz [1973, 1976a]).

Purpose of the Present Study

The researcher wishing to select a method for detecting item bias is confronted with many methods and no clear guidelines for choosing among them. Thus Rudner (1976), and Scheuneman (1977) have noted the need to compare the various methods in order to determine which are best. If all the approaches were to validly identify the same biased items, one could use the simplest, least expensive approach. However, if the approaches isolate different items, it is necessary to identify those methods which produce the most valid results.

Thus, the purpose of this study is to compare and contrast several of these item bias methods so that guidelines may be provided for their use. Four bias methods, two of which can be easily implemented from classical item analysis techniques, and two of which stem from a more sophisticated attempt to control for ability differences using item characteristic curves will be applied to actual test data. In addition two indices intended as validity measures, will be applied. The research questions to be examined include:

1. To what extent do the four methods agree or disagree in the identification of biased items?
2. Which methods are in highest agreement with the validity measures?
3. What is the content of items identified as biased?

REVIEW OF RELEVANT BIAS TECHNIQUES

For the sake of brevity, only those methods of direct relevance to the present study will be reviewed.

Bias Method Based on Transformed Difficulty

A number of approaches have used item difficulty (proportion of correct responses) as the focus of analysis. An item is considered biased in this approach if, compared to other items on the test, it is relatively more difficult for one group than for another. One of the more widely implemented techniques of this type is that used by Angoff (1972). It will be referred to as the transformed difficulty method. (A similar procedure was also suggested by Thurstone (1925) in connection with his method of absolute scaling.) The method involves computing the difficulty or p-value (proportion of subjects getting the item right) for each item separately for each group. In order to obtain an interval scale of item difficulty values, each p-value is transformed to a normal deviate or z-value from standard tables of the normal distribution. Then to avoid negative z-values, a delta value is calculated from the z-value by the equation $\Delta = 4z + 13$. A large delta value indicates a difficult item.

For two groups, there will be a pair of delta values for each item. These pairs of delta values can then be plotted on a graph, each item represented by a point on the graph. Figure 1 illustrates this procedure. The plot of these points appears similar to a correlation diagram and, in fact, the correlation between the delta values is used to describe the plot, i.e. to

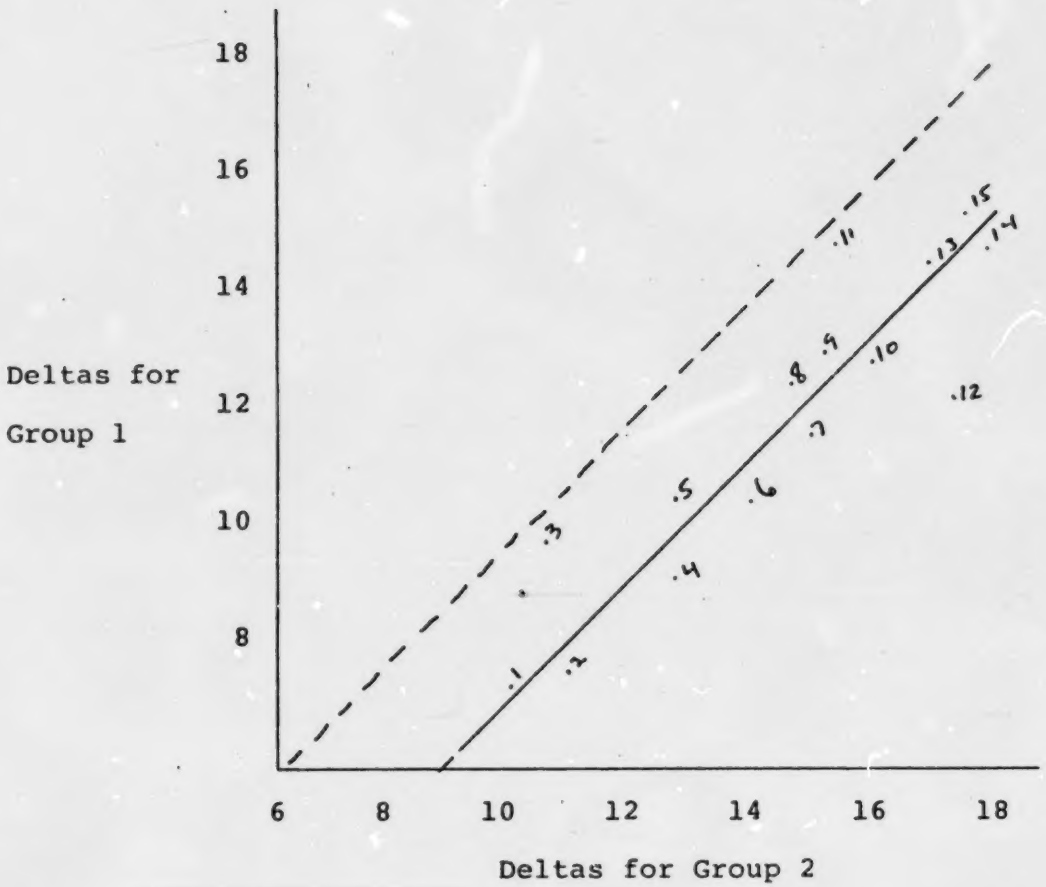

Figure 1. A hypothetical bivariate plot of delta values for two groups illustrating Angoff's method of determining item bias.

describe the performance similarity or difference of two groups. As in Figure 1, this correlation is typically high, indicating that, in general, the rank order of the delta values (or item difficulties) is similar in both groups. Furthermore, as illustrated by the solid line in Figure 1, a line can be fitted to the plot of points; and the deviation of a given point from the line is taken as measure of that item's bias, large deviations indicating much bias. Thus, in Figure 1, items 3, 11, and 12 are all biased. Notice that all items fall below the equal delta line (dotted). This indicates that the test as a whole is more difficult for Group 2. It is also interesting to note that although items 3, 11, and 12 are all more difficult for Group 2, only item 12 would be considered biased against Group 2. Items 3 and 11 would be biased against Group 1, since they are relatively more difficult for that group.

The transformed difficulty procedure has been used to study cultural differences in a wide variety of contexts (Gulliksen, 1964; Angoff, 1975, Angoff and Sharon, 1972; Angoff and Modu, 1973; Angoff and Ford, 1973; Breland et. al., 1974; Rudner, 1976). Since the primary purpose of this study is methodological rather than substantive, content findings are not discussed but the interested reader is advised to consult the above references.

Bias Methods Based on Item Discrimination

Another type of statistic that has been used in the study of bias is the item discrimination index. Traditionally, the correlation between an individual item and the total test score

has been used in standard item analyses to identify discriminating items. More recently attempts have been made to use this index to identify biased items.

Green and Draper (1972) were among the early investigators of the point biserial approach to item bias. They investigated differences in responses to several sections of the 1970 California Achievement Tests for groups of northern, urban blacks; southern, rural blacks; southwestern Mexican-Americans; and advantaged whites. Items were considered biased if they were in the best discriminating half of the items for one group and in the worst half for another group. The investigators found the median proportion of biased items across all pairs of groups to be .30. Sustantive interpretations of the biased items were not given. A check was made which insured that the findings were not just due to random variation. Since the number of biased items was substantial, Green and Draper concluded that using the discrimination index as the basis for item selection, tests built with minority group response data would differ by about 20-50% in items chosen as compared to tests built with majority group data.

The Chi-Square Method

Instead of focusing on a single item parameter, like difficulty or discrimination, other methods compare entire distributions of responses for the two groups in question. Scheuneman's chi-square procedure is one such technique.

According to her definition, (1975):

> ...An item would be considered unbiased if for persons

> with the same ability in the area being measured, the probability of a correct response on the item is the same regardless of the population group membership of the individual (p. 2).

This definition assumes the subtest score is a reasonably valid measure of ability, as indicated below. Operationally, the definition may be restated:

> An item is unbiased if, for all individuals having the same score on a homogeneous subtest containing the item, the proportion of individuals getting the item correct is the same for each population group being considered (p. 2).

Thus, each race is divided into various ability groups based on their observed subtest scores, and within each ability group the proportion of correct responses is compared between races. The statistic used for this purpose is a modified chi-square.

Intuitively, then, this method detects as biased an item in which the p-values within each score level are different for the different ethnic groups. Note that the overall p-values (total proportion correct) between groups may be very different and yet the item may not be biased according to this definition.

Scheuneman's chi-square method has been applied primarily to the Metropolitan Readiness Test (MRT) given to kindergarten and beginning first graders. In a study of the item pool used for the development of the 1976 MRT, responses were analyzed for black/white differences (Scheuneman 1975, 1977). Eighty out

of 579 or 14% of the items were identified as biased. In research done on the 1975 standardization of the MRT, 34 out of 97 items were found to be biased across five groups--White, Black, Mexican-American, Puerto Rican and Oriental (Scheuneman, 1976). Clear, interpretable causes were identified for 14 of these items.

The Item Characteristic Curve Method

As illustrated in Figure 2, an item characteristic curve describes the relationship between an examinee's Z-score on tested ability and the probability of his answering an item correctly. Birnbaum (1968) formalized a three parameter curve model to describe this relationship as discussed in Lord and Novick (1968). The curve which relates the probability (p) of a correct response to item g for a person of ability θ_i is given by:

$$P = C_g + \frac{1 - C_g}{1 + e^{-1.7a_g(\theta_i - b_g)}}$$

where a_g is an item discrimination index, b_g is an item difficulty index, and C_g is a guessing parameter.

Figure 2 illustrates how parameters a_g, b_g and c_g are related to the curve. The difficulty parameter b_g indicates the position of the curve's inflection point along the Z-continuum. More precisely, if $c_g = 0$ the value of b_g is the ability z-score for a person who has a .50 probability of getting the item right. The larger the value of b_g, the more difficult the item. A value related to the slope of the curve at b_g is a_g, the discrimination

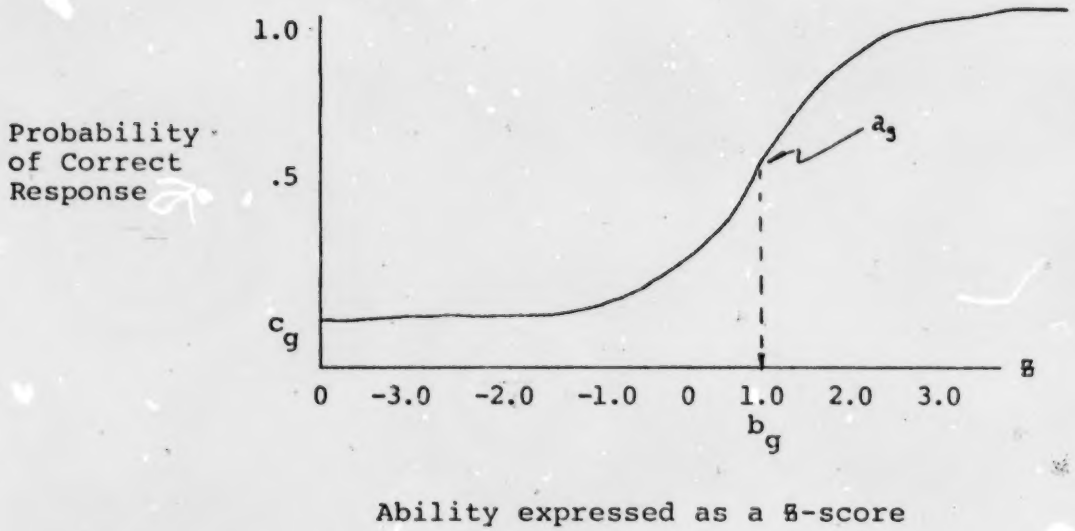

Figure 2. A hypothetical Item Characteristic Curve.

parameter. For larger values of a_g the slope is steeper, and thus the item is better able to distinguish examinees of close ability. Finally, c_g, the guessing parameter, is the probability of correct response for an examinee of very low ability.

The LOGIST procedure (Wood, Wingersky & Lord, 1976) used in this study involves estimating the a_g, b_g, and c_g parameters that determine the curve for each item g. This is done separately for each ethnic group of interest. An unbiased item is one whose a_g, b_g, and c_g values are the same, after equating, for different ethnic groups.

In an exploratory study involving approximately 2,250 whites and the same number of blacks, Lord (1977) analyzed an 85 item verbal test (the Verbal section of the April, 1975 SAT) consisting of four kinds of items--verbal analogies, antonyms, word meaning, and reading comprehension. Forty-six of the 85 items were found to be biased.

Rationale for Choice of Methods

In the design of this study, an effort was made to select the four item bias methods (discussed above) which are the major focus of current use and/or research. Due to practical limitations, it was not possible to compare all item bias methods.

The transformed difficulty procedure and the discrimination differences procedure are both readily implementable from standard item analysis output. Thus, the cost of implementation and difficulty of interpretation is minimal. However, these are the lease mathematically sophisticated procedures and, as will be

explained later, are most subject to variations in the particular subject sample studied. At the other end of the continuum is the ICC procedure developed by Lord which uses the most sophisticated and current knowledge in the field. This method is mathematically most precise and has the great advantage of providing estimates of the difficulty and discrimination parameters which are independent of the ability of the particular sample of examinees tested. The problem with Lord's approach comes when considering the ease of implementation. The method requires a large number of subjects (roughly 1,000), is very expensive, and requires a fairly extensive mathematical background to understand the technique.

The chi-square method is conceptually similar to the ICC method. It attempts to determine if people of the same ability in different groups have the same probability of getting the correct answer. However, Scheuneman's method is easier for the lay person to understand, is less costly and requires fewer subjects.

Other Studies of Item Bias

Rudner (1977) recently completed a dissertation in which he compared an item difficulties procedure, a factor score approach, Scheuneman's chi-square technique and an item characteristic curve (ICC) approach to 48 items from a reading comprehension subtest. He applied these techniques to (a) two diverse culture group samples (normal and hearing impaired) and (b) two equal culture group samples (two groups of hearing impaired with differing mean scores). Of most relevance to the present study

were the correlations he found of .67 between the chi-square and ICC approach, .59 between the chi-square and item difficulties approach, and .31 between the ICC and transformed item difficulties approaches. Rudner concluded that although there was some agreement between the ICC and chi-square approaches and the transformed item difficulties and chi-square approaches, the agreement was not overwhelming.

Nurgester (1977) recently completed a dissertation in which he compared three bias techniques introduced by Angoff, Fishbein and Scheuneman respectively. He was looking at male/female bias on English (54 items) and mathematics (73 items) tests. Of most relevance to this study are the obtained correlations of .65 and .92 between the Angoff and Scheuneman approaches on the English and mathematics tests respectively.

METHOD

The Sample

The sample used in the present study was drawn from 17,726 twelfth grade students tested as part of the National Longitudinal Study (NLS) of 1972. Sampling procedures as well as other details are in Hilton and Rhett (1973). The original sample of 17,726 students was reduced to 14,828 by the requirement that, for NLS purposes each student had to have both an NLS Student Questionnaire asking for background information and an NLS Student Test Book answer sheet. The ethnic compositon of the final sample of 14,828 subjects was as follows: American Indian-178; Black-1,895; Spanish-Oriented (Mexican-American, Puerto Rican, Other Latin)-677;

Oriental-176; White Northeastern-2,798; White North Central-3,589; White Southern-3,557; White Western-1,958.

The Instruments

Subtests were taken from the parent version of the NLS battery. The composition of the battery was determined by the objective of obtaining a comprehensive description of people with diverse backgrounds, ethnic affiliations, and socio-economic status. The sequential order of the subtests was designed to alternate three traditional with three non-traditional tests as follows: Vocabulary, Picture-Number, Reading, Letter Groups, Mathematics, and Mosaic Comparisons.

Each of the tests was designed to tap a different cognitive skill. The Vocabulary test consisted of 15 items using synonym format. The Picture-Number test involved an associative memory task whereby students study a page of pictures paired with numbers and are subsequently asked to recall the number associated with each picture. The reading test is composed of questions based on short passages. The passages were taken from the Project Access Reading Test and deal with topics of particular relevance to minority students. An item on the Letter Groups test consists of five groups of letters. The student must identify which group is different from the rest. Thus, it is designed as a test of inductive reasoning. The Mathematics test items require the student to compare two quantities and state whether they are equal, unequal (and which is greater) or whether the answer is not ascertainable from the information provided. The final test

in the battery, Mosaic Comparisons, is a speeded test which measures perceptual speed and accuracy. Subjects must locate, for each of a pair of mosaics, the column that contains a single square which is shaded differently.

Procedure

Data from the six subtests of the NLS Study of 1972 were analyzed using the four previously described methods of detecting item bias (transformed difficulty, discrimination differences, chi-square, and item characteristic curve) as applied to white and black ethnic groups.[1]

Bias Methods

Transformed difficulty method. Angoff's delta plot procedure as previously described was applied. Each item had a delta value for whites (x) and for blacks (y). Then a cross-plot of delta values was created as in Figure 1. The major axis line of the ellipse of points is given by the linear equation $y=ax+b$, where

$$a=\frac{(s_y^2-s_x^2)+\sqrt{(s_y^2-s_x^2)+4r_{xy}^2s_x^2s_y^2}}{2r_{xy}s_xs_y}$$

and

$$b=M_y-aM_x.$$

The symbols s_x, s_y, r_{xy}, M_x, and M_y represent the standard

---

[1]Due to cost restrictions, only half of the White North Central group (N = 1,794) was analyzed together with the total black group (N = 1,895). The second half of the white North Central group was used for extra analyses described later in the paper.

deviation of x and y, correlation between x and y, and mean of x and y respectively.

The distance of item point i from the major axis is given by:

$$d_i = \frac{ax_i - y_i + b}{\sqrt{a^2 + 1}}$$

where x_i and y_i are the white and black delta values associated with item i. This distance is the absolute value measure of bias for the transformed difficulty approach. The larger d_i is in absolute value, the more biased the item. A signed transformed difficulty measure of bias, which preserved both the direction and magnitude of bias was obtained by attaching a positive sign to d_i if the item was relatively easy for whites and a negative sign if the item was relatively difficult for whites.

Discrimination differences method. An approach analogous to the one used by Green (1972) was applied. For each item, the point biserial correlation of item and total test scores was computed separately for blacks and whites. One index of bias was the absolute value of the difference between the white and black point biserial coefficients. Again, the larger this index, the more biased the item. The direction of bias could be preserved by attaching the appropriate sign to the difference. This signed bias index was the white minus black point biserial difference.

Chi-square method. Scheuneman's chi-square procedure was applied next. The χ^2 statistic is the index of bias, large values

indicating much bias. Only magnitude and not direction of bias was preserved for this index.

Table 1 illustrates how the analysis is performed. The first step is to divide the subtest score into a number of ability levels (usually three to five). This is generally done by dividing the distribution of correct responses on the item for the smaller population into quartiles or quintiles and then using the same cutoffs for both populations. In Table 1, the second set of columns contains the frequency distributions of blacks, whites, and the combined group. Following that, the third set of columns gives the number who got the item right for each score range and within each ethnic group. The proportion correct is arrived at by dividing the total number who got the item right by the total number whose subtest score falls within the given range. This proportion is then multiplied in the final columns by the number within each ethnic group to arrive at the expected frequencies (e.g., .9116 x 12 = 10.94). The same proportion is used in arriving at the expected frequencies for both groups because, according to the definition, if an item is unbiased, the proportion of correct responses should be the same regardless of group membership.

The chi-square statistic is then computed from the expected and obtained frequencies by using the formula at the bottom of Table 1. This procedure differs from the conventional chi-square test in two fundamental ways: (a) the computation of the expected frequencies and (b) the calculation of the degrees of freedom.

Table 1

Example of Chi-Square Computation for One Item[a]

Total Score on Subtest	Number with Scores in Each Range			Number Answering Item Correct			Proportion Correct	Expected Frequencies	
	Black	White	Total	Black	White	Total	(T_o/T)	Black	White
	B	W	T	B_o	W_o	T_o	p	B_e	W_e
12-18	12	350	362	10	320	330	.9116	10.94	319.06
10-11	34	152	186	17	104	121	.6505	22.12	98.88
8-9	24	66	90	8	40	48	.5333	12.80	35.20
3-7	33	47	80	6	15	21	.2625	8.66	12.34

$$\chi^2 = \Sigma \frac{(B_e - B_o)^2}{B_e} + \Sigma \frac{(W_e - W_o)^2}{W_e} = 5.317$$

[a]Adapted from Scheuneman, 1976

While the conventional degrees of freedom would be (r-1)(k-1), the modified degrees of freedom used here is (r-2)(k-1) where r is the number of score levels and k is the number of ethnic groups. Finally, the significance of the chi-square statistic can then be found and the determination of biased items made. Note that this procedure is easily generalized to more than two groups.

Item characteristic curve (ICC) method. Finally, the item characteristic curve method described previously was applied. For each item, the LOGIST computer program (Wood & Lord, 1976; Wood, Wingersky & Lord, 1976) was used to estimate the a and b parameters separately for blacks and whites. These parameter estimates were equated by computing the between group linear regression for the b parameters and using the reciprocal of the slope of that line for transforming the a parameter. The c parameter in each case was fixed at the value of $.75(1/N)^2$ where N is the number of multiple choice options.

The absolute value measure of bias was the area between the white and black item characteristic curves. The absolute value area bias index was calculated by adding successive rectangles whose width was .005 and whose height was the absolute value of the difference between the probability of getting the item right at each particular ability level. This procedure was recommended and used by Rudner (1977b).

---

[2]Suggested by Lord (personal communication).

The signed item characteristic curve method proceeded by adding successive rectangles described above, but the height for each rectangle was determined by the white minus the black probability of getting the item correct at each particular ability level. Thus when the white curve was above the black curve the area for a rectangle was added to the bias index and when the black curve was above the white, the rectangles were subtracted. A negative area would indicate an item biased against whites.

Research Questions

Agreement Among Bias Methods

For each of the four methods (transformed difficulty, discrimination differences, chi-square and ICC) described in the previous section, absolute value bias measures were computed for each test item. Pearson product-moment correlations were then computed across items for each pair of methods, both within each of the five subtests and between all items in the battery as a whole. As another measure of similarity, the 24 most biased items for each method were examined and the percent of agreement or overlap was computed for each pair of methods.

Further analyses compared the agreement among the three bias methods for which both direction and magnitude of bias could be maintained. This was done by computing Pearson r's both within subtests and between all items in the battery as a whole.

Validity of Bias Methods

The four bias indices were also correlated with two types of validity measures.

One pervasive question in bias research is whether differences are due to ability or to bias. It might be argued that an analysis should be based on white and black groups matched on ability (Lord, 1977). The basic premise is that a search for bias is one in which we seek to find items functioning differently for persons of the same ability in different groups. While some of the methods attempt to control for ability differences, artifacts are still introduced because the distributions of ability are markedly different in the two groups to be compared. Hunter (1975) examines how this may cause unbiased items to be identified as biased. Rudner (1977) shows how the chi-square technique may become inflated as the total observed score distributions differ.

One way to insure that the ability distributions are the same is to match groups on ability. For each subtest taken, a random sample of 100 blacks was chosen and 100 whites were matched on the basis of subtest total score to the black sample. To minimize problems associated with matching on the same test and population to be studied, the white group used for matching was from the randomly selected second half of the white north central group and was not used elsewhere in the study. The index used for measuring bias of each item was either the signed white minus black "p" value difference for matched groups, or the absolute value of the subtest matched p difference.

The second validity measure was based upon classifying each item as traditional (verbal and mathematics) or nontraditional (letter groups, mosaic comparisons, picture number). It

has been argued that traditional items are more closely associated with the cultural experiences of the white middle class. Therefore, one might expect more of these items to be biased. To test this, the bias index for each method was correlated across item with the traditional-nontraditional (one/zero) classification.

Content of Biased Items

The content of biased items was examined for possible insight into the sources of bias. This was done at a descriptive level.

RESULTS

Descriptive Information

After excluding whites and blacks with missing data, the sample sizes for the analysis were 1691 blacks and 1794 whites. Summary statistics on the relative performance of blacks and whites across subtests are given in Table 2.

In general, the performance of whites was consistently higher than blacks. The difference varied from 4.4 to 8.0 points and was not noticeably greater for the traditional verbal and mathematical tests as compared to the non-traditional tests. In fact the largest mean difference occurred on the mosaic comparisons test. However there was a considerable overlap in performance of the two groups as evidenced by the size of the standard deviations relative to the mean differences. Except for the mosaic comparisons test, the variances were quite similar for black and white groups. The reliabilities of the subtests were also quite similar for both groups.

Table 3 extends the descriptive information to the item level.

Table 2

Subtest Means, Standard Deviations and Reliabilities

for Two Ethnic Groups

		Subtest				
		Verbal	Picture Number	Letter Groups	Math	Mosaic Comparisons
Statistic	Group	35 items	30 items	25 items	25 items	40 items
Mean	White	20.0	17.7	18.3	16.5	32.8
	Black	12.8	13.3	13.1	10.3	24.8
Standard Deviation	White	6.4	7.0	4.4	5.3	6.8
	Black	5.7	6.8	5.1	4.7	9.2
Hoyt Reliability	White	.85	.88	.83	.85	.93
	Black	.81	.88	.84	.78	.94

Table 3

Subtest Item Statistics

Item Statistic	Group	Subtest[a] Verbal 33 items	Picture Number 30 items	Letter Groups 24 items	Math 23 items	Mosaic Comparisons 40 items
p-Value	White	.59(.18)	.61(.11)	.75(.16)	.68(.13)	.82(.20)
	Black	.37(.19)	.43(.14)	.54(.20)	.42(.15)	.61(.26)
Point Biserials	White	.37(.08)	.54(.07)	.44(.09)	.45(.09)	.52(.13)
	Black	.40(.08)	.52(.05)	.46(.10)	.39(.08)	.55(.15)

[a]Numbers in parenthesis represent standard deviations.

Five of the 155 item composites were excluded from this and subsequent analyses.[3]

Bias Information

Each of the bias indices were calculated separately within each subtest. Table 4 summarizes this information. The transformed difficulty, chi-square and ICC approaches all have their highest bias indices for the verbal subtest. Chi-square and ICC both have the second highest bias values for the math subtest. Finally, the discrimination difference index seems to be unrelated to the traditional non-traditional classification of the subtests.

Agreement Among Bias Methods

To facilitate the measurement of agreement among the bias methods, the intercorrelation matrices of Table 5 were calculated. The correlations for the entire battery of 150 items are rather low indicating that the agreement among the methods is rather poor. The largest correlation was between the chi-square and ICC approach. This was not totally surprising since these are based on the same theoretical rationale. Other significant correlations ($\alpha = .01$) were found between the chi-square and the transformed difficulty, and the transformed difficulty and ICC approach. The discrimination differences approach did not correlate significantly with any other method.

---

[3]For one of these items, the parameter estimates did not stabilize. The other four were excluded because the p-value for examinees not removed from LOGIST was less than .75 (1/N), where N is the number of options.

Table 4

Average Bias Indices by Subtest[a]

Subtest	UNSIGNED BIAS METHODS[b]			
	Transformed Difficulty	Discrimination Differences	Chi-Square	ICC
Verbal	.475 (.239)	.076 (.069)	8.21 (6.91)	.354 (.158)
Picture-Number	.235 (.198)	.037 (.332)	5.49 (7.06)	.169 (.102)
Letter Groups	.278 (.177)	.080 (.069)	4.62 (3.90)	.283 (.219)
Mathematics	.311 (.246)	.082 (.070)	7.05 (5.75)	.327 (.185)
Mosaic Comparisons	.334 (.246)	.081 (.039)	3.44 (3.40)	.178 (.081)
Overall	.331 (.238)	.071 (.058)	5.64 (5.78)	.255 (.167)

[a]Numbers in parenthesis represent standard deviations.

[b]In this table the bias methods maintain only magnitude and not direction of bias.

Table 5

Correlations of Unsigned Item Bias Indices

for Entire Battery and by Subtest

		Unsigned Bias Methods			
		1	2	3	4
Subtest	Method	Transformed Difficulty	Discrimination Differences	Chi-Square	ICC
Entire	2	-.091	---		
Battery	3	.370**	.154	---	
(150 items)	4	.239**	.137	.485**	---
Verbal	2	-.206	---		
(33 items)	3	.414*	.351*	---	
	4	.373*	-.149	.361*	---
Picture-	2	.213	---		
Number	3	.501*	.075	---	
(30 items)	4	.265	.252	.652**	---
Letter	2	-.421*	---		
Groups	3	.206	-.007	---	
(24 items)	4	-.161	.171	.742**	---
Math	2	-.097	---		
(23 items)	3	.138	.264	---	
	4	.245	.269	.505**	---
Mosaic	2	-.196	---		
Comparisons	3	.401**	-.058	---	
(40 items)	4	.088	-.100	-.047	---

**Significant at $\alpha = .01$
*Significant at $\alpha = .05$

From Table 5 it is also possible to examine the consistency of the correlations between methods across subtests. There was considerable variability in the size of correlations across subtests. In particular, it appears as if the low chi-square, ICC correlation on the mosaic comparisons subtest is aberrant. Lord (personal communication) has said the ICC method is not appropriate for speeded tests. If this correlation is removed from the analysis, the overall agreement between the chi-square and ICC methods would be even higher.

As another measure of agreement, Table 6 presents the percent agreement for the 24 most biased items under each method. Twenty-four was chosen somewhat arbitrarily, since not all of the methods have significance tests associated with them. Significance tests for the chi-square and matched group methods indicated 24 might be a reasonable number to consider.

From Table 6 it is evident that the agreement among the methods in general follows the pattern of correlations. By chance alone, one expects 16% agreement. This is arrived at by multiplying the probability of identification of an item as biased (24/150 x 24/150 x 150 = 3.84) by any two procedures by the number of items, and converting this to a percent (3.84/24 = 16%). The only pairs of methods which differed substantially from the 16% agreement expected by chance alone were the chi-square with ICC, transformed difficulty with ICC and transformed difficulty with chi-square. From a practical standpoint, the agreement was rather low.

Table 6

Percent Agreement Among Bias Methods for 24 Most Biased Items*

	Bias Indices			
	Transformed Difficulty	Discrimination Differences	Chi-Square	I.C.C. Area
1. Angoff's Delta	—			
2. Point Biserial Differences	8.3	—		
3. Chi-Square	37.5	20.8	—	
4. I.C.C. Area	33.3	20.8	54.2	—

*Expect 16% agreement by chance alone.

A reanalysis of the intercorrelations between methods for those methods where both direction and magnitude of bias could be preserved was performed and is presented in Table 7. This analysis was performed because it was felt that one was losing information by not preserving direction of bias and thus the correlations in Table 5 would underrepresent the true agreement among the methods. As can be seen in Table 7, preserving the direction of bias had the effect of substantially increasing the correlation between the transformed difficulty and ICC approaches, but no effect on the correlation of the discrimination differences procedure with the others.

Comparison of Bias Indices, Matched Group Differences and Traditional-Nontraditional Classification of Items

The correlations between the bias indices and the matched group "p" differences and traditional-nontraditional classification of items are presented in Table 8. Although the magnitude of these correlations was small, significant correlations were found between the traditional-nontraditional classification and the ICC, chi-square and transformed difficulty approaches in that order.

Moderately small but statistically significant correlations were also found for the ICC and transformed difficulty approaches with the matched group index. A closer look within subtests in Table 9 reveals that these correlations are substantially higher within some of the subtests, but fall apart for the Mosaic Comparisons subtest.

Table 7

Correlations of Signed Bias Indices

for Entire Battery and by Subtest

		Signed Bias Methods		
Subtest	Method	1 Transformed Difficulty	2 Discrimination Differences	3 ICC
Entire Battery (150 items)	2	.080		
	3	.491**	.079	-
Verbal (33 items)	2	.407*		
	3	.687**	.236	-
Picture-Number (30 items)	2	-.309		
	3	.675**	-.184	-
Letter Groups (24 items)	2	.155		
	3	.318	.362	-
Math (23 items)	2	-.106		
	3	.419*	-.037	-
Mosaic Comparisons (40 items)	2	-.127		
	3	.364*	-.147	

**Significant at $\alpha = .01$
*Significant at $\alpha = .05$

Table 8

Correlations of Validity Measures with Item Bias Indices for the Entire Battery

Bias Indices	Validity Measures	
	Matched Group[a]	Traditional-Nontraditional[b]
Transformed Difficulty	.279*	.244*
Discrimination Differences	-.030	.100
Chi-Square	---	.280*
ICC Area	.316*	.414*

*Significant at = .01 level

[a]The matched group index (white minus black p for equal ability groups matched on subtest scores) was correlated with the signed bias indices.

[b]The traditional-nontraditional index was correlated with the unsigned bias indices since maintaining signs on the bias indices for this analysis would have made the correlations meaningless.

Table 9

Correlation of Subtest Matched Group Index with Bias Indices by Subtest

Subtest	Signed Bias Indices		
	Transformed Difficulty	Discrimination Differences	ICC
Verbal	.420*	.260	.310
Picture Number	.580**	.047	.264
Letter Groups	.102	.090	.251
Math	.407*	.158	.605**
Mosaic Comparisons	-.163	-.534**	.002

**Significant at $\alpha = .01$
*Significant at $\alpha = .05$

Content of Biased Items

Next, an examination of those items listed as the ten most biased by each method was made. Although it was difficult to find the source of bias in many of these items, some interpretable trends were apparent. First, there was a noticeable preponderance of verbal items. Second, items at the end of the subtests were frequently identified as biased. Third, two problems which were biased in favor of blacks were an item from a reading passage dealing with black television and an item on the math subtest dealing with money. As compared to the rest of the test these may be more commonly encountered in the every day lives of blacks. Finally, two items dealing with temperature were biased against blacks. One was a math problem involving degrees Fahrenheit and the other was a picture of a thermometer paired with a number for a memory task.

DISCUSSION

Agreement Among Four Bias Methods

For 150 items across five subtests the correlations among three of the four bias methods examined, the transformed difficulty, chi-square and item characteristic curve approaches were statistically significant although small in magnitude . The point biserial differences did not significantly correlate with any of the methods. Both the correlations and percent agreement suggested that chi-square and ICC indexes were most similar followed by the transformed difficulty procedure. An examination of the correlations among methods for the various subtests

revealed a considerable amount of variability in agreement. Further investigation of the correlations was afforded by comparing several methods' signed bias indices (which preserved the direction as well as magnitude of bias). The correlation between the transformed difficulty and ICC methods increased substantially. One might also hypothesize an increase in the chi-square, ICC correlation increase if the sign could be maintained for the chi-square.

Possible Reasons for Disagreement

To ascertain the explanation of the disagreement, a closer look was taken at those items identified as biased by one method but not another. For this comparison, an item was biased by one method and not another if, within a subtest, the bias index for an item was above 1 standard deviation above the mean for one procedure and below 1 standard deviation below the mean for another procedure.

Rudner (1977) found that all the items identified as biased by both the transformed difficulty and the item characteristic curve procedures were items where the discrimination parameters were similar, but the location parameters differed. He referred to this as non-inverted differential performance and it is illustrated in Figure 3a. The present study confirmed this as one type of item biased by both procedures, however, items which criss-crossed also fell in this category. For equal ability or unequal ability groups, one would expect relatively large p-differences (and thus delta differences) for items of the type

Rudner found. However with groups of unequal observed distributions it is possible for items of type 3b to be biased by both. The area is quite large and the p-differences are also large. Recall that the p-differences are derived from observed distributions and not ability distributions. The illustrations are exaggerated to make the point clear.

Items biased by the ICC approach but not by the transformed difficulty approach were of the type pictured in 3c. They crisscrossed but the p-differences were of a magnitude similar to the average p-differences for that subtest.

Parts d and e illustrate those items identified as biased by the transformed difficulty procedure but not by the ICC approach. Even when the ICC curves are identical in two groups, the p-differences can be relatively small or large because of differences in the observed score distribution.

The comparison of Angoff to Scheuneman follows similar logic.

The point-biserial differences approach identified items whose ICC curves crossed or, in the case of Scheuneman's chi-square, whose p values changed direction as the total test score intervals increased. Thus the area could be substantial and the r's could be similar as would be the case in Figure 3a.

Given that the ICC and chi-square approach follow the same logic, it was surprising to find that the correlation was not higher. It seemed however that the results from the mosaic comparisons subtest was pulling that correlation down. Since

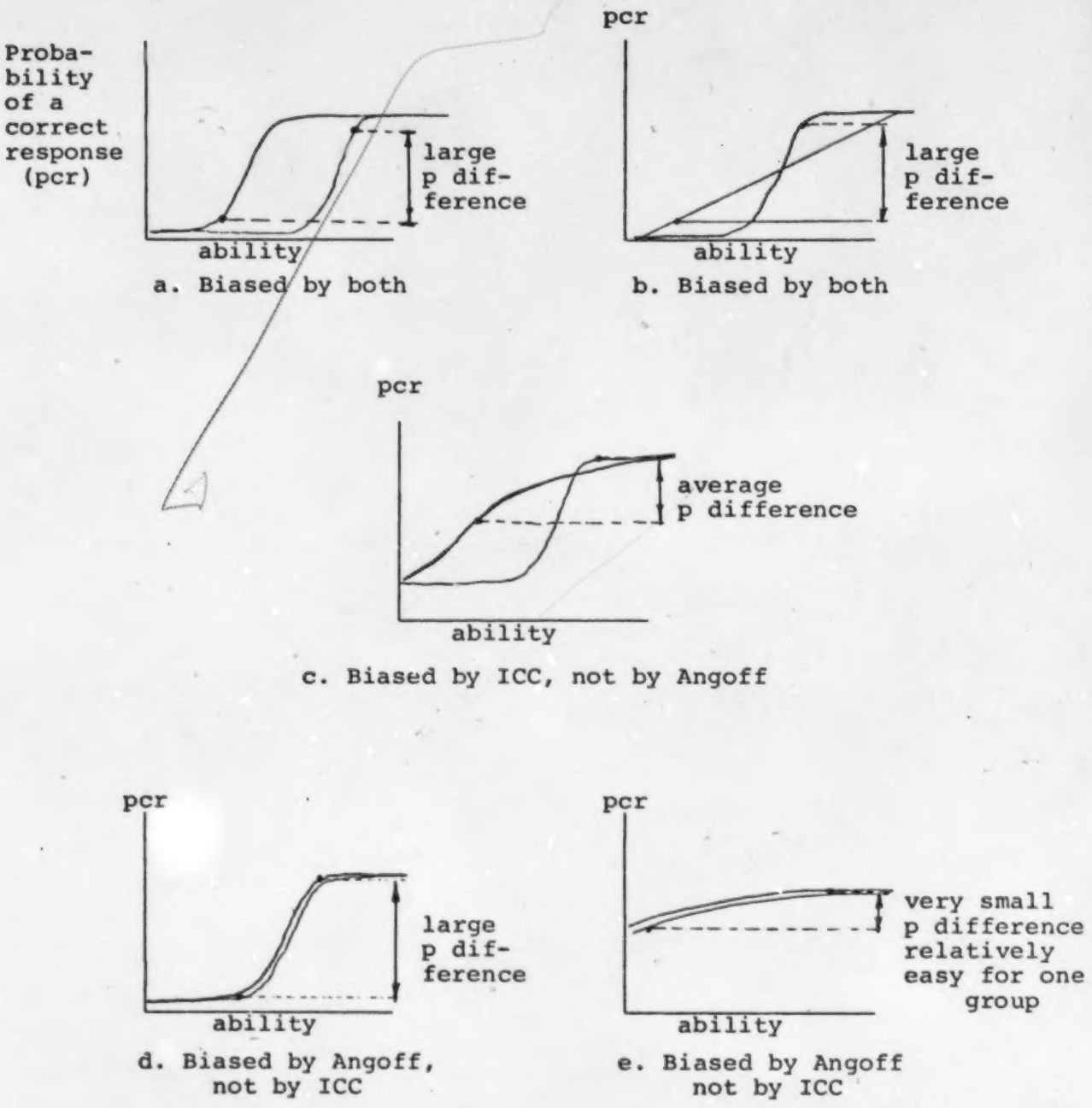

Figure 3. Various items identified as biased or not by Angoff delta and ICC approaches.[a]

[a]The "p" values for two groups are indicated by dots on the diagram. Observed distributions are separate but overlapping.

the ICC approach was not appropriate with that subtest, the observed correlation should have been somewhat higher.

Although the logic is the same for both the ICC and chi-square methods there are several differences which could explain a discrepancy in bias. These are (1) the total test score is used to estimate ability directly in the chi-square approach and only indirectly in the latent trait approach, (2) the chi-square approach collapses the ability continuum into intervals whereas the ICC approach maintains this information, (3) the ICC approach assumes a model relating ability to probability of a correct response which may not be correct for all items.

Other theoretical reasons for disagreement in the identification of biased items are given by Hunter (1975). He discusses several factors which may cause an item to be labeled as biased when in fact no bias exists. These are (1) non unidimensional tests, (2) differences in ability distribution of the two groups, (3) differences in item quality, (4) guessing, and (5) nonlinearity of regression. He showed how an unbiased test may have item by race interaction. Specifically items of different difficulty will show an item by group interaction in any situation in which the two groups differ in ability. Angoff's procedure is additionally criticized because it does not take into account item quality and distortion produced by guessing. Hunter also claims that differences in point biserial coefficients may be a result of comparing groups with different mean scores rather than as a result of bias. This is because the point biserial correlation

is largest if an item has a p value of .50. Next, several flaws in the item characteristic curve method are noted. Differences in the ability distributions will be reflected in instability at the ends of the curves for different groups, and different displacements of the item characteristic curves because of unreliability. He finally points out that all the methods fail if the test is not unidimensional.

Bias in the Test Used

Another possibility for the generally low correlations among methods may have been that the test used for the analysis was not very biased. Therefore, the methods may have been correlating essentially random fluctuations of the bias indices. However, two analyses fail to support this possibility.

First, the half of the white group not used in this study was compared to the half that was used. If there were no bias in the test we would expect the discrimination difference to be the same for the white versus white and black versus white comparisons. In addition, if there were no bias in the test we would expect the variability of the bias measures (distance of deltas to the axis, discrimination difference) to be the same for both black versus white and white versus white comparison group.

Reference to Table 10 shows that the discrimination differences are larger for the black versus white comparison, that these differences have higher variability in the black versus white comparison, and that the variability for the distance measure

Table 10

A Comparison of Several Statistics in the Same and Different Culture Groups

Statistic	Culture Comparison[1]	Subtest: Verbal	Picture Number	Letter Groups	Math	Mosaic Comparisons
Discrimination Difference[2]	WW	.026	.022	.029	.020	.025
	BW	.079	.037	.084	.089	.081
Distance Measure Variability[3]	WW	.10	.12	.12	.12	.16
	BW	.54	.31	.33	.42	.42
Discrimination Difference Variability[4] (in hundredths)	WW	.03	.02	.05	.03	.05
	BW	.52	.11	.50	.57	.15

[1]WW = white versus white culture comparison.
BW = black versus white culture comparison.

[2]Absolute value of black minus white (or white minus white) point biserial calculated for each item and then averaged.

[3]Standard deviation of distance measure (with sign maintained) used in transformed difficulty approach.

[4]Variance of discrimination difference described above. This is reported in hundredths so .52 really is .0052.

used is also greater for the black versus white comparison rather than the white versus white comparison. All of this suggests there is much more than just error involved in the black versus white comparison.

Additional evidence for bias in the test comes from the significance test associated with the chi-square method. For a significance level of .05, 36 of the 150 or 24% of the items were biased in the black versus white comparison. In contrast, only 3 or 2% of the items were found to be biased in the white versus white comparison.

Both analyses suggest there is a substantial amount of bias in the battery.

Agreement Across Subtests

It is important to note that while the correlations among the four bias methods were generally not high and the agreement on the items identified as most biased was not strong in all cases, there was substantial agreement (Table 4, p. 27) in terms of identification of the bias in different subtests. For example, the transformed difficulty, chi-square and ICC approaches all suggest that the traditional verbal and mathematics tests (as a group) contain more bias than the three non-traditional subtests.

That the differences in bias among subtests is not simply a reflection of differences in performance can be seen by the consistency of differences in Table 3.

The bias indices may have a tendency to be greater on less

homogeneous subtests. Although the math and verbal tests of Table 3 are less homogeneous (have lower reliability) these differences are so small it is unlikely that homogeniety would have been a major factor underlying bias differences among subtests.

Comparison to Other Studies

As was noted earlier, Rudner (1977b) has also completed a dissertation on item bias. He utilized a 48 item verbal test and compared normal and hearing impaired subjects. He found correlations of .67 between the chi-square and ICC approach, .59 between the chi-square and transformed difficulty approach, and .31 between the ICC and transformed difficulty (all unsigned) approaches. These were substantially higher than the analogous correlations of .49, .37 and .24 found in this study, although the ordering was the same.

Several reasons for this are possible: (1) Rudner only used one subtest. In the present study there was considerable variability in the size of the correlation coefficients across subtests; they were as high as .74, .50, .37 respectively for the pairs of methods mentioned above. These values more closely approximate his results.

(2) His population of subjects, normal versus hearing impaired, is very different from a black versus white culture group comparison.

(3) More bias may have been present in the test he used. He found 56% of the items (Rudner, 1977b) (p. 39) to be biased

according to the chi-square method ($\alpha = .001$). This study found 24% of the items to be biased according to the chi-square method at $\alpha = .05$.

(4) The transformed difficulty and ICC procedures were slightly different from the ones used in this study. Rudner transformed the p values to within group normal deviates and later used the equal difficulty line for his transformed difficulty procedure. Also in the ICC approach, he used a different estimation procedure, did not fix the guessing parameter, and found the area from -5σ to $+5\sigma$.

A direct comparison cannot be made to Nungester's (1977) study. Although two of the methods he compared were used in this study (Angoff's transformed difficulty and Scheuneman's chi-square) in order to use an index which preserved direction, he dropped many of the items from the chi-square analysis.

Traditional-Nontraditional Classification of Items

Three of the methods correlated significantly with a classification of items as traditional (verbal, math) and nontraditional (picture-number, mosaic comparisons, and letter groups). This meets with intuitive expectations, and based on the validity of the assumption that verbal and math tests should exhibit more bias because they appear to be more closely associated with the experiences of the dominant culture, seems to offer some support for the validity of three of the measures. In particular the ICC method correlated highest, followed by the chi-square and transformed difficulty methods.

Subtest Matched Group "p" Differences

The subtest matched group index was calculated by matching on subtest total score a white sample (not used elsewhere in the study) with a random sample of blacks for each subtest. For each item, a "p" value or difficulty value was calculated in the black and in the white sample. The white minus black "p" became the signed index; the unsigned index was the absolute value of this number.

It may be argued that this index should be regarded as another bias index only and not as a validity index. The rationale for its introduction is that as a result of the sampling both ethnic groups will have the same ability distribution. Other bias methods attempt to control for ability, but as was noted earlier, artifacts are introduced into each of the procedures because the ability distributions are not equal. Whether one regards it as another way of measuring bias or as a validity index, it is still useful to find its convergent validity with the other methods.

Evidence was found for significant correlations between this method, the transformed difficulty, and ICC methods with the direction of bias maintained on all. The chi-square method was not included in this analysis.

Limitations of the Study

The current study uses only two culture groups and a set of items for which the amount of bias is unknown a priori. It assumes both biased and unbiased items are present.

Although a large number of subjects can compensate for a small number of items, the number of items used to obtain the parameter estimates was on the low side (5 to 40). It should be noted that the parameter estimates did converge on all subtests although they were not appropriate for the speeded test.

Some specific procedures used may limit the generality of the study. The area was used as a measure of similarity between item characteristic curves rather than a significance test. In addition the area was taken from -3 to +3 σ. The regression line was used for equating rather than the major axis.

Future studies should consider external validity measures as well.

Summary and Conclusions

Test data from two diverse culture groups (1,691 blacks; 1,794 whites) were analyzed to determine the agreement among four methods of bias and two indices intended as validity criteria. The test battery was given as part of the National Longitudinal Study of High School Class of 1972 and was composed of 155 items from six subtests: vocabulary, reading comprehension, math, letter groups, picture-number, and mosaic comparisons.

There appeared to be some bias in the battery as measured by (1) those indices having associated significance tests, and (2) a comparison of two white groups showing considerably less bias than the black and white comparison groups of interest.

The four bias methods examined were transformed difficulty

(Angoff, 1972), discrimination differences, Scheuneman's (1975) chi-square, and an item characteristic curve (ICC) approach. (Lord, 1977; Rudner, 1977b.) For the 150 items analyzed, three of the methods (which maintained magnitude but not direction of bias), (1) transformed difficulty, (2) chi-square, and (3) ICC approaches were moderately correlated (1 & 2 = .239, 1 & 3 = .370, 2 & 3 = .485). There was little agreement between the discrimination differences approach and the others. Upon close examination the percent agreement among the items identified as most biased by each method was moderately low.

Of the two validity indices, the traditional versus nontraditional classification of items had higher correlations with the bias methods; the obtained correlations were (.244; .100; .280; and .414) with the four bias methods respectively. The second validity measure examined difficulty differences for a random sample of 100 blacks matched with 100 whites separately on each subtest score. This index correlated significantly with both the transformed difficulty (.279) and ICC (.316) approaches (with both direction and magnitude of bias preserved). The chi-square index was not included in this analysis because the direction of bias is difficult to maintain.

Finally, the correlations were not high enough to offer conclusive advice to the user on which bias methods to employ. Thus, additional studies need to be done to further evaluate the external validity and comparability of the methods.

REFERENCES

Angoff, W. H. A technique for the investigation of cultural differences. Paper presented at the Annual Meeting of the American Psychological Association, Honolulu, May 1972.

Angoff, W. H. The investigation of test bias in the absence of an outside criterion. Paper presented at the NIE Conference on Test Bias, 1975.

Angoff, W. H. General considerations in the study of test bias when the criterion is absent. Paper presented at the Annual Meeting of the American Educational Research Association, April, 1976.

Angoff, W. H. & Ford, S. F. Item-Race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 1973, 10, 95-105.

Angoff, W. H. & Modu, C. C. Equating the scales of the Spanish language Prueba de Aptitude Académica and the English language Scholastic Aptitude Test. Research Bulletin 73-4; CEEB RDR 72-73, No. 4 Princeton, N.J.: Educational Testing Service, 1973.

Angoff, W. H. & Sharon, A. L. Patterns of test and item difficulty for six foreign language groups on the Test of English as a Foreign Language. Research Bulletin 72-2; CEEB RDR 71-72, No. 5, Princeton, N.J.: Educational Testing Service, 1972.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical Theories of Mental Test Scores. Reading, M.A.: Addison-Wesley, 1968, Chapters 17 through 20.

Breland, H. M., Stocking, M., Pinchak, B. M. & Abrams, N. The cross-cultural stability of mental test items: An investigation of response patterns for ten socio-cultural groups. Project Report 74-2. Princeton, N.J.: Educational Testing Service, 1974.

Burrill, L. E. An investigation into the nature of ethnic bias in items and tests assessing current educational status. Unpublished dissertation proposal. Fordham University, N.Y., 1975.

Cardall, C. & Coffman, W. E. A method for comparing the performance of different groups on the items of a test. Research Bulletin 64-61, Princeton, N.J.: Educational Testing Service, 1964.

Cleary, T. A. & Hilton, T. L. An investigation into item bias. Educational and Psychological Measurement, 1968, 8, 61-75.

Cronbach, L. J. Five decades of public controversy over mental testing. American Psychologist, 1975, 30, 1-14.

Darlington, R. B. A Defense of "rational" personnel selection and two new methods. Journal of Educational Measurement, 1976, 13, 31-41.

Echternacht, G. A quick method for determining test bias. Educational and Psychological Measurement, 1974, 34, 271-280.

Green, D. R. & Draper, J. F. Exploratory studies of bias in achievement tests. A paper presented at the Annual Meeting of the American Psychological Association, Honolulu, September, 1972.

Green, D. R. Racial and Ethnic Bias in Test Construction. Monterey: CTB/McGraw-Hill, 1972.

Green, D. R. Racial and Ethnic Bias in Achievement Tests and What to do About it. Monterey, CA: CTB/McGraw-Hill, 1973.

Green, D. R. Reducing bias in achievement tests. A paper presented at the annual meeting of the American Educational Research Association, San Francisco, April, 1976.

Gulliksen, H. Intercultural attitude comparisons and introductory remarks at the Princeton University Conference on Preference analysis and Subjective Measurement. Research Memorandum 60-8, Princeton, N.J.: Educational Testing Service, 1964.

Hays, W. L. Statistics for the Social Sciences, Holt, Rinehart and Winston, Inc., N.Y. 1973.

Hilton, T. L. & Rhett, H. The Base-Year Survey of the National Longitudinal Study of the High School Class of 1972. Final Report, Contract No. OEC-0-72-0903, Office of Education, National Center for Educational Statistics, U. S. Department of Health, Educationa and Welfare. ETS Princeton, N.J. June, 1973.

Hunter, J. E. A critical analysis of the use of item means and item-test correlations to determine the presence or absence of content bias in achievement test items. A paper presented at the NIE conference on test bias; Maryland; December, 1975.

Jensen, A. P. An examination of culture bias in the Wonderlic Personnel Test. Arlington, VA: Eric Clearinghouse, 1973, (ERIC Document Reproduction Service ED 086 726).

Johnson, S. C. Hierarchical clustering schemes. Psychometrika, 1967, 32, 241-54.

Lord, F. M. A study of item bias using item characteristic curve theory. Preliminary report, ETS, 1977.

Lord, F. M. & Novick, M. R. Statistical Theories of Mental Test Scores, Addison-Wesley, Mass., 1968.

Merz, W. R. Factor analysis as a technique in analyzing item bias. Paper presented at the annual meeting of the California Educational Research Association, Los Angeles, CA 1973.

Merz, W. R. Estimating bias in test items utilizing principal components analysis and the general linear solution. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA, 1976a.

Merz, W. R. Test fairness and test bias: A review of procedures. A paper presented at the Office of Education Invitational Conference on Achievement Testing of Disadvantaged and Minority students for Educational Program Evaluation, Reston, VA, May, 1976b.

Nungester, R. An empirical examination of three models of item bias. Unpublished doctoral dissertation, Florida State University, 1977.

Ozenne, D. G., Van Gelder, N. C. & Cohen, A. J. Emergency school aid act (ESAA) national evaluation, achievement test standardization. Santa Monica, CA: Systems Development Corporation, 1974.

Pothoff, R. F. Statistical aspects of the problems of biases in psychological tests. University of North Carolina Institute Statistical Mimeo Series, 1966, No. 479.

Rudner, L. M. Item and format bias and appropriateness, unpublished paper, 1976.

Rudner, L. M. An approach to biased item identification using latent trait measurement theory. Paper presented at the annual meeting of the American Educational Research Association, New York, April 1977a.

Rudner, L. M. An evaluation of select approaches for biased item identification. Unpublished doctoral dissertation, Catholic University of America, 1977b.

Rudner, L. M. Approaches toward biased item identification, submitted for publication, 1977c.

Scheuneman, J. A new method of assessing bias in test items. Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C. April, 1975.

Scheuneman, J. A procedure for evaluating item bias in the absence of an outside criterion. A paper presented at the annual meeting of the American Educational Research Association, San Francisco, April, 1976.

Scheuneman, J. A method of assessing bias in test items in the absence of an outside criterion, preliminary draft, The Psychological Corporation, 1977.

Swineford, F. Test analysis: National longitudinal study of the high school class of 1972, ETS, SR-73-29, 1973.

Thorndike, R. L. Concepts of cultural-fairness. Journal of Educational Measurement, 1971, 8, 63-70.

Thurstone, L. L. A method of scaling psychological and educational tests. Journal of Educational Psychology, 1925, 16, 433-451.

Veale, J. R. & Forman, D. I. Cultural variation in criterion referenced tests: A global item analysis. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April, 1976.

Williams, R. L. Black pride, academic relevance, and individual achievement. The Counseling Psychologist, 1970, 2, 18-22.

Williams, R. L. Abuses and misuses of testing Black children. The Counseling Psychologist, 1971, 2, 62-73.

Wood, R. L. & Lord, F. M. A user's guide to LOGIST. Research Memorandum, Educational Testing Service, Princeton, New Jersey, 1976.

Wood, R. L., Wingersky, M. S. & Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters, Research Memorandum, ETS, 1976.